AD_____

GRANT NUMBER: DAMD17-94-J-4174

TITLE: Utilizing Serial Measures of Breast Cancer Risk Factors

PRINCIPAL INVESTIGATOR(S): Mimi Y. Kim, Sc.D.

CONTRACTING ORGANIZATION: New York University Medical Center
New York, New York 10010-2598

REPORT DATE: January 1996

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 1996	Annual (31 Dec 94 - 30 Dec 95)

4. TITLE AND SUBTITLE
Utilizing Serial Measures of Breast Cancer Risk Factors

5. FUNDING NUMBERS
DAMD17-94-J-4174

6. AUTHOR(S)
Mimi Y. Kim, Sc.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
New York University Medical Center
New York, New York 10010-2598

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick
Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Failure to account for errors-in-measurement of exposure and confounder variables can result in biased estimates of relative risk. We have developed a technique for correcting for measurement error when subjects have a variable number of repeated measurements and the average of the measurements is used as the subject's measure of exposure in the analysis. A bootstrap method for obtaining confidence intervals of the relative risk estimates, which takes into account the variability in the estimates of the correction factors, has also been devised. In addition, we have considered a model for adjusting hormone measurements for systematic fluctuations over the menstrual cycle, which will allow more valid comparisons of hormone levels between pre-menopausal cases and controls. Finally, most studies on the half-lives of environmental contaminants such as PCBs, which have been linked with an increased risk of breast cancer, have been based on small sample sizes and a limited number of repeated measurements. We have developed guidelines for choosing the number of subjects, repeated measurements, and time interval between measurements in the design of half-life studies.

14. SUBJECT TERMS	15. NUMBER OF PAGES
breast cancer biostatistics epidemiology	52
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- **DOD** - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- **DOE** - See authorities.
- **NASA** - See Handbook NHB 2200.2.
- **NTIS** - Leave blank.

Block 12b. Distribution Code.

- **DOD** - Leave blank.
- **DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- **NASA** - Leave blank.
- **NTIS** - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

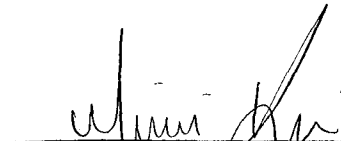
____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.



PI - Signature

1/11/96

Date

TABLE OF CONTENTS

Overview

With prospective cohort studies incorporating the periodic collection of blood samples from participants, one has the greater potential to study the temporal relationship of biomarkers and other risk factors to the development of disease than with retrospective case-control studies. However, statistical methods for analyzing repeated measurements have not been fully explored and developed. Furthermore, methods for estimating and correcting for errors-in- measurement using repeated measurements remain obscure and are seldomly applied in epidemiologic studies. The objectives of this project are to apply as well as develop theoretical statistical methods for utilizing repeat determinations of serum levels of endogenous hormones and other biologic measurements in the analysis of nested case-control studies of breast cancer, and to estimate and correct for errors-in-measurement.

This progress report describing research accomplished during the first year of the grant period is comprised of three chapters. In Chapter I, we describe a technique for correcting for measurement error when subjects have a variable number of repeated measurements, and the average of the measurements is used as the subject's measure of exposure in the analysis. Failure to account for errors-in-measurement of exposure and confounder variables can result in biased estimates of relative risk, obscuring the true relationship between an exposure variable with breast cancer risk. Our method applies a correction factor to the subject's average value, prior to model fitting, which is a function of the number of repeated measurements. The resulting logistic regression estimate based on the corrected exposure measurement is unbiased. A bootstrap method for obtaining confidence intervals, which takes into account the uncertainty in the reliability estimates is also proposed. A manuscript based on this work is currently in preparation.

In Chapter II, we describe a method for adjusting for the systematic variability of hormone levels over the menstrual cycle, based on a mixed ANOVA model with cubic splines. The method standardizes hormone measurements obtained at different time during the menstrual cycle, thus allowing for more valid comparisons of hormone levels between premenopausal breast cancer cases and controls. This research is still in progress.

Finally, in order to fully elucidate the role of environmental contaminants, such as PCBs, in the development of breast cancer, their rates of persistence in the body must be accurately quantified. Individuals who are able to clear the toxic compounds from the body at a faster rate (as measured by the half-life of the toxin) may be at lower risk of breast cancer. Published estimates of the half-life of PCBs, however, have been widely variable, ranging from .5 months to 17 years. The lack of consistency among study estimates may be largely due to the small sample sizes and limited number of repeated measurement per subject utilized in these studies. Guidelines for choosing the number of repeats and the optimal time interval between repeats for estimating an individual's half-life with a given level of precision, while minimizing the cost of the study, have been developed and are described in Chapter III. Furthermore, sample size and power considerations for studies comparing two-population half-lives are also presented. A paper describing this work has been accepted for publication by Archives of Environmental Contamination and Toxicology.

Chapter I

Correcting for Measurement Error in the Analysis of Case-Control Data with Repeated Measurements of Exposure

1 Introduction

In most case-control studies, the risk factors of interest are measured with error. For biologic variables, such as blood pressure, nutrient, and hormone levels, measurement error can arise from limitations in the measurement technique or laboratory assay. In addition, because the exposure of interest is usually a subject's underlying long-term average value rather than the level at any single point in time, intrinsic fluctuations in the variable over time can also contribute to measurement error.

When the error is random and non-differential with respect to case-control status, it is well known that estimates of relative risk based on the mis-measured exposure will be attenuated. In order to minimize the effects of measurement error, many investigators advocate collecting repeated measurements of the exposure on all subjects and using the individual's average value (De Klerk et al., 1989). However, as noted by Rosner et al (1992), even when the mean of several replicates is substituted for a single measurement, attenuation of relative risk may still occur, especially when the average is based on only a few repeats or when the degree of measurement error is large.

Methods for correcting estimates of relative risk for measurement error have been addressed in a number of epidemiologic and statistical papers (Armstrong et al., 1989; Thomas et al., 1993). The most common method involves correcting the "naive" relative risk estimate based on the observed exposure by the expected amount of bias. In the case of logistic relative risk regression, the regression parameter will be attenuated by the factor, R, which is equal to the reliability coefficient of the mis-measured exposure (Rosner et al., 1992, De Klerk et al., 1989). Therefore, one can multiply the biased estimate of the regression coefficient by the inverse of the reliability coefficient to obtain the corrected estimate. This method, however, is dependent on the assumption that the reliability of the exposure measurement is the same for all subjects. When the average of several replicates is used as the measure of exposure, this condition will be met only if all subjects have an equal number

of repeated measurements, given the degree of measurement error associated with a single measurement is the same for all subjects.

In studies in which the exposure is measured on repeated occasions, however, subjects often have a variable number of measurements because of missing data. For example, the data that are utilized to illustrate the methods in this paper are derived from the NYU Women's Health Study, a nested case-control study of serum hormonal levels and breast cancer (Toniolo et al). The study cohort consists of 15,785 women who donated multiple blood samples over time and have been followed since enrollment for the development of breast cancer. Most women have donated one or two samples; however, many have also donated three or more.

Because subjects with a larger number of multiple blood samples have a more precise measure of their true underlying serum hormonal levels than those with fewer measurements, the reliability of the average of the available measurements will correspondingly vary between subjects. Consequently, if the observed average is used as the measure of exposure for each subject, the usual procedure for correcting for measurement error cannot be applied.

Liu and Liang (1992) proposed an estimating equation approach for obtaining consistent estimates of logistic regression parameters when all subjects have the same number of repeated imprecise exposure measurements, which in principle could be extended to the more complicated situation when the number of replicates is variable between subjects. In this paper, we discuss an alternative method for correcting for measurement error in the analysis of matched case-control data when subjects have a variable number of repeated exposure measurements and the individual's average is used as the measure of exposure. A bootstrap algorithm for obtaining confidence intervals, which takes into account the variability due to estimation of the reliability coefficent is also proposed. The methods are illustrated using data from a nested case-control study of estradiol levels and risk of breast cancer from the NYU Women's Health Study.

2 Methods

In describing the methods below, we assume the measurement error model of Armstrong et al (1989) for matched case-control studies. The techniques are generalizable to the unmatched design by assuming there is only one matching stratum.

Let x_{ij} denote the true value of the exposure variable for the j^{th} subject in stratum i, for $i = 1, ..., M; j = 1, ..., s_i$ Assume that x_{ij} is normally distributed with mean, $\mu_i + \delta$ if the subject is a case, or μ_i if she is a control, and variance σ_s^2. In addition, let z_{ijk} denote the k^{th} repeated observation of x_{ij}, for $k = 1, ..., n_{ij}$. Then, assuming the classical errors-in-variables model, we have:

$$z_{ijk} = x_{ij} + e_{ijk},$$

where the error term, e_{ijk}, is independent of x_{ij} and $e_{ijk'}$ for $k \neq k'$, and normally distributed with mean 0 and variance, σ_e^2. It follows that the observed z_{ijk} in stratum i are normally distributed with means $\mu_i + \delta$ and μ_i for cases and controls, respectively, and common variance, $\sigma_s^2 + \sigma_e^2$. The variance component, σ_s^2, can be interpreted as the between-subject variance of the true exposure, adjusted for matching stratum and case/control status, and σ_e^2 as the variance due to measurement error.

Under these assumptions, Armstrong et al (1989) showed that the probability that a study subject is a case, conditional on an observed average based on n measurements, $\bar{z}$, and membership in stratum i, is a logistic function:

$$\Pr(D = 1|\bar{z}; i) = \frac{\exp(\alpha_i + \beta R_n \bar{z})}{1 + \exp(\alpha_i + \beta R_n \bar{z})}, \tag{1}$$

where

$$R_n = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2/n} \tag{2}$$

is the reliability of $\bar{z}$ as a measure of x. When no measurement error is present, $\bar{z} = x$, the reliability coefficient is equal to 1, and (1) reduces to:

$$\Pr(D = 1|x; i) = \frac{\exp(\alpha_i + \beta x)}{1 + \exp(\alpha_i + \beta x)}.$$

Thus, an estimate of the logistic regression coefficient based on $\bar{z}$ will estimate the "naive" coefficient, $\beta^* = \beta R_n$, rather than the true β. Because the reliability coefficient is between 0 and 1, the "naive" β^* will be attenuated relative to β. We can see from (2), however, that as the number of repeated measurements increases, the reliability coefficient approaches 1, and the corresponding attenuation in β will diminish.

When all subjects have the same number of n repeated measurements, an unbiased estimate of the regression coefficient can be obtained by fitting the logistic model using $\bar{z}$ for each subject's exposure measurement, and multiplying the resulting coefficient estimate, $\hat{\beta}^*$, by $1/R_n$. If subjects have a variable number of measurements, however, this approach cannot be applied, since the reliability of the exposure variable would no longer be constant for all subjects, but would depend on the number of available repeated measurements.

For the case where the reliability of the exposure differs across subjects, the regression coefficient may be obtained by correcting a subject's average exposure measurement by the relevant reliability coefficient, prior to model fitting. That is, if the j^{th} subject in stratum i has the observed average $\bar{z}_{ij.}$, based on n_{ij} approximate measurements of x_{ij}, then replacing the unknown x_{ij} in the conditional logistic model with the transformed average, $R_{n_{ij}}\bar{z}_{ij.}$, where $R_{n_{ij}}$ is calculated from (2), will yield an unbiased estimate of β. Since the reliability is higher for larger n_{ij}, this method effectively gives more weight to the averages based on a large number of repeats and less weight to those based on few repeats.

This method for correcting for errors-in-measurement is analogous to the "two-stage" approach discussed in Thomas et al. (1993) in which the expected value of the true exposure given the data, $E(x_{ij}|\bar{z}_{ij.})$, is computed and then used as the exposure in the usual conditional logistic regression model. Whittemore (1989) and Prentice (1982) have proposed similar methods for correcting for errors-in-variables in linear and Cox proportional hazards regression models, respectively.

Although fitting the logistic model to the transformed covariate will result in an unbiased estimate of β, the corresponding variance of $\hat{\beta}$ will be underestimated unless the reliability

coefficient is known. Usually, however, the variance components in (2) must be estimated from a separate reliability substudy or from the subset of subjects in the main study with repeated measurements.

Assuming we have N_r subjects with replicate measurements, we can estimate the variance components, σ_s^2 and σ_e^2, as follows. Let $\mathbf{z}_i^* = \{z_{i1}^*, ..., z_{ik_i}^*\}$ denote the k_i repeated observations of the exposure for the i^{th} subject in the reliability study sample. Then one can estimate σ_e^2 by calculating

$$\hat{\sigma}_e^2 = \sum_{i=1}^{N_r} \sum_{j=1}^{k_i} (z_{ij}^* - \bar{z}_{i.}^*)^2 / \{(\sum_{i=1}^{N_r} k_i) - N_r\}. \tag{4}$$

For reasons of efficiency, σ_s^2 should be estimated from all subjects in the main study. Because the total within-stratum variance, σ_T^2, is equal to $\sigma_s^2 + \sigma_e^2$, the between-subjects variance can be estimated by subtracting $\hat{\sigma}_e^2$ from $\hat{\sigma}_T^2$, which may be obtained using the first measurement of each subject in the main study and fitting the model:

$$z_{ij} = \mu_i + \delta c_{ij} + e_{ij},$$

where μ_i denotes the overall mean for stratum i, c_{ij} denotes the case ($c_{ij} = 1$) or control ($c_{ij} = 0$) status for the j^{th} subject in the i^{th} matched set, and e_{ij} is the residual error. The mean-squared error from the above model will estimate σ_T^2. Then, $\hat{\sigma}_s^2$ can be calculated from $\hat{\sigma}_T^2 - \hat{\sigma}_e^2$. Given $\hat{\sigma}_s^2$ and $\hat{\sigma}_e^2$, it follows that $R_{n_{ij}}$ can be estimated as $\hat{\sigma}_s^2/(\hat{\sigma}_s^2 + \hat{\sigma}_e^2/n_{ij})$.

When R_n is estimated, variances and confidence intervals for $\hat{\beta}$ based on the transformed covariate must take into account the extra variability due to estimation of the reliability coefficient. Rosner et al. (1992) have derived the asymptotic variance of the corrected logistic regression parameter, which includes the uncertainty of the reliability estimate, for use in cohort studies under a rare disease assumption. Their method, however, is applicable only when all subjects in the main study have the same number of repeats.

For the situation when subjects in a case-control study have a variable number of replicates, we propose the following bootstrap procedure for obtaining confidence intervals for

the corrected $\hat{\beta}$. We assume that among the N subjects in the main study, the N_r subjects with at least 2 measurements are used for the reliability study data.

1. Generate a bootstrap sample from the reliability data. Consider the vector of observations, $\mathbf{z}_{ij} = \{z_{ij1}, ..., z_{ijn_{ij}}\}$, from a particular subject as the sampling unit. In order to keep the total number of repeated measurements in each bootstrap sample the same for all iterations, utilize a "stratified" sampling scheme where N_2 observation vectors are sampled with replacement from the N_2 subjects with 2 repeats, N_3 observations from the subjects with 3 repeats, and so on.

2. Estimate σ_e^2 from the reliability bootstrap sample using (4).

3. Generate a bootstrap sample from the main study data, where the sampling unit is the matched set. If the number of subjects in each stratum differs across strata, utilize a stratified scheme analogous to the above to keep the total number of subjects constant. That is, sample M_2 matched sets from the M_2 sets with 2 subjects, M_3 sets from the matched sets with 3 subjects, etc...

4. Using the main bootstrap sample, estimate σ_t^2 and σ_s^2.

5. For each subject in the main bootstrap sample, transform the subject's observed average by multiplying by the appropriate correction factor from (2).

6. Estimate β by fitting a conditional logistic regression model, with x_{ij} replaced by the transformed covariate for all subjects.

Repeat (1-6) 1,000 times to generate the bootstrap distribution of β, which is the approximate minimum number of bootstraps necessary to compute bias-corrected confidence limits (Efron and Tibshirani, 1986). The simple $(1-\alpha)\%$ confidence interval can be constructed using the $\alpha/2$ and $(1-\alpha/2)$ percentiles of the bootstrap distribution. Bias-corrected confidence intervals should be used when the bootstrap distribution of β is asymmetric and when

the sample sizes of the main and reproducibility studies are small (Efron and Tibshirani, 1986). We report only the bias-corrected confidence intervals in this paper.

Extensions to Multicovariate Models

Thus far, our focus has been on correcting for measurement error in a single exposure variable, in the absence of confounders. However, the methods can also be generalized to the multi-covariate situation, where the confounders, in addition to the primary exposure variable, may be measured with error. We give a brief outline of the methods below, but refer the reader to Armstrong et al (1989) for additional details on the measurement error model and estimation of variance components.

In order to generalize the techniques to the multivariate situation, assume that $\mathbf{x}_{ij}$ denotes a $(p \times 1)$ vector of true covariates for the i^{th} subject in stratum j, and that it follows a multivariate normal distribution with mean vector $\mu_i + \Delta$ for the cases and μ_i for the controls, and covariance matrix Σ. In addition, let

$$\mathbf{z}_{ijk} = \mathbf{x}_{ij} + \mathbf{e}_{ijk}$$

denote the k^{th} observed measurement of $\mathbf{x}_{ij}$, for $k = 1, ...n_{ij}$, where the $\mathbf{e}_{ijk}$ are independent and identically distributed according to a multivariate normal distribution with covariance matrix, Ω.

Under these assumptions, Armstrong et al (1989) showed that the probability a subject is a case, conditional on the mean of n observed replicate covariate vectors, $\{\mathbf{z}_1, ..., \mathbf{z}_n\}$, is equal to the following logistic function:

$$Pr(D = 1|\bar{\mathbf{z}}_\cdot, i) = \frac{\exp(\alpha_i + \bar{\mathbf{z}}_\cdot \Lambda_n \beta)}{1 + \exp(\alpha_i + \mathbf{z}_\cdot \Lambda_n \beta)},$$

where $\bar{\mathbf{z}}_\cdot = (\sum_{k=1}^{n} \mathbf{z}_k)/n$, $\Lambda_n = (\Sigma + n^{-1}\Omega)^{-1}\Sigma$, and β is the $(p \times 1)$ vector of logistic regression parameters.

When subjects have a variable number of replicate measures of the exposure variables, it follows that as in the single covariate case, one can transform the observed mean covariate vector for each subject by multiplying the vector by an estimate of the matrix, $\Lambda_{n_{ij}}$, and then fit the usual logistic regression model to the transformed covariates to obtain the corrected logistic regression coefficients for all covariates. A bootstrap algorithm analogous to that for the single covariate case could be used to obtain corrected confidence intervals which take into account the variation due to estimation of $\Lambda_{n_{ij}}$, but the method could become very computationally intensive with a large number of confounders, since more complicated multivariate MANOVA models would be needed to estimate Σ and Ω. For the special case when the confounders are measured without error, however, estimation of the variance components is greatly simplified (see Kim et al (1995)), and the bootstrap method could be more easily applied.

3 Example

The primary aim of the NYU Women's Health Study is to determine whether serum levels of endogenous hormones, such as estradiol, are associated with risk of breast cancer. Between March 1985 and June 1991, a cohort of healthy women aged 34-65 years were enrolled at the Guttman Breast Diagnostic Institute, New York. At the time of enrollment and at annual screening visits thereafter, women were asked to donate blood and complete a self-administered questionnaire. Serum samples were frozen and stored for future biological assays. Subsequent cases of breast cancer were identified primarily through active follow-up and confirmed by reviewing medical and pathological records. In this example, only the women who were post-menopausal at enrollment are included.

In order to limit the costs associated with measuring hormone levels in the cohort, a nested case-control study design is used. For each incident case of breast cancer, individually matched controls are selected at random from the risk set consisting of all cohort members

alive and free of breast cancer at the time of diagnosis of the case, and who match the case on menopausal status at entry, age at entry, and number and approximate dates of blood donations up to the date of diagnosis in the case. For additional details of the study design, see Toniolo et al (1991).

The goal of this example is to evaluate the effect of random measurement error on the associations between total, % free, and % bound to sex hormone binding globulin (SHBG-bound) estradiol levels and risk of breast cancer, when the average of all the available repeated measurements for a subject is used as her exposure. The associations between the baseline measurements of the total, % free, and % SHBG-bound estradiol levels and risk of breast cancer among post-menopausal women, unadjusted for measurement error, were evaluated by Toniolo et al (1994). Total and % free estradiol were found to be positively associated with risk of breast cancer, whereas % SHGB-bound estradiol had a strong protective effect.

Using data from both post-menopausal cases and controls, we estimated the reliability coefficients for total, % free and % bound estradiol, adjusted for matching stratum and case/control status, as: .47, .67, and .91, respectively (Table 1). (These estimates were somewhat lower than those published by Toniolo et al (199): .51, .77, and .94 for total, % free and % bound estradiol, respectively, which were based on data from only the post-menopausal controls in the NYUWHS.) The estimates of the reliability coefficients indicate that the degree of measurement error in total and % free estradiol may be sufficiently large to attenuate observed relationships with risk of breast cancer.

The main case-control study sample consisted of 379 subjects stratified into 130 matched sets. Ten matched sets had 1 control per case, 119 sets had 2 controls per case, and one set had 3 controls per case. Of the 379 subjects in the main study, the 157 (41%) with 2 or more repeated measurements were used for the reproducibility data set. Ninety-eight subjects had 2 replicates, 53 had 3 replicates, and 6 subjects had 4. Estradiol values were log-transformed to improve normality, and to be consistent with the scale used in the main study.

We investigated the effects of measurement error on the observed associations between each exposure variable and risk of breast cancer by comparing the estimated logistic regression parameters based on the first measurement of the exposure for each subject, the average of the replicate measures, and the transformed (corrected) average value. Corresponding odds ratios were calculated from the regression estimates by comparing women in the 90^{th} versus 10^{th} percentiles of the observed distributions (i.e., 76.5 vs 14.5 for total estradiol, 1.7 vs. 1.04 for % free, and 57.6 vs 27.3 for % SHBG-bound estradiol).

Bootstrap confidence intervals were generated by utilizing the SAS macro facility in conjuction with PROC PHREG, a procedure which can be used for fitting conditional logistic regression models. All analyses were run on a DEC 3000/700 AXP computer workstation.

The results are provided in Table 2. For total estradiol and % free estradiol, the uncorrected analyses show that using the average of the repeated measurements results in a minor increase in the regression coefficient compared with using only the baseline measurement. On the other hand, the regression coefficients corrected for measurement error based on the transformed averages are substantially larger than the uncorrected estimates for both variables.

The effect of measurement error on the estimated odds ratios is especially striking. When comparing women in the 90th percentile versus the 10th percentile of the observed total estradiol distribution, the corrected odds ratio was estimated to be 9.70, compared with uncorrected odds ratios of 3.02 and 3.60 using the baseline and untransformed average, respectively. Similiarly, the corrected odds ratio for % free estradiol was 5.10, compared with 3.07 for the baseline measurement and 3.13 for the average value.

This illustrates how using the observed average of replicate measurements of exposure for each subject may not be sufficient to offset the effects of measurement error when the degree of error is large and when subjects have only a few replicates, and that additional error correction procedures may be necessary. In the case of total estradiol, one would need to take the average of 10 replicate measurements to improve the reliability to .90, based

on the estimated variance components in Table 1. For % free estradiol, one would need 5 measurements. Thus, it is not surprising that using the average value in our example did not appreciably deattenuate the corresponding regression coefficient since less than half the study population had 2 or more measurements. On the other hand, because % SHGB-bound estradiol levels are highly reproducible, the logistic regression estimates and corresponding odds ratios using the corrected average were not very different from the uncorrected analyses.

As one would expect, the bias-corrected boostrap confidence intervals shown in Table 2 for the true regression coefficient are shifted further away from 0 and are wider than the uncorrected confidence intervals, since the bootstrap method accounts for the variation due to estimation of the variance components in the reliability coefficient.

4 Conclusions

Haukka (1995) proposed a similar bootstrap method for correcting for measurement error in generalized linear models for the situation when the "gold standard" is known for the exposure measurement and validation, as opposed to reproducibility, data are available. When compared with the correction method for logistic regression proposed by Rosner et al (1990) which also takes into account the variability in $\hat{R}$, the bootstrap method was found to yield wider confidence intervals for peaked and skewed measurement error distributions. As discussed by Haukka (1995), this difference may result because the bootstrap method takes better account of the measurement error variance, whereas the Rosner et al. method is based on a first-order Taylor series approximation, which may not adequately correct confidence intervals when the error variance is large.

In utilizing the average value of repeated measurements as the exposure, one must assume that the individual measurements are distributed randomly around the unobserved true value, and that levels of the exposure are not changing systematically over time. Among breast cancer cases, however, hormone levels could be influenced by the development of dis-

ease so that measurements obtained closer to the date of diagnosis may exhibit a systematic time trend. Preliminary analyses using linear regression techniques, however, suggest the absence of any trends in estradiol levels over time for the cases (results not shown).

The error correction method proposed in the paper is also dependent on the assumptions that the true exposure and error are normally distributed, with variances σ_s^2 and σ_e^2, respectively, which are homogeneous across strata and case/control status. Because only one case is included in each stratum, we could not evaluate whether σ_s^2 is constant for cases and controls. However, σ_e^2 for total estradiol was estimated as .16 and .18 for cases and controls, respectively, suggesting the error variances are similar.

The distributions of % free and % SHBG-bound did not deviate significantly from normality, so no transformations were necessary. Total estradiol, on the other hand, had a skewed distribution so the data were log-transformed to improve normality.

We have shown that in situations when the magnitude of measurement error is large and subjects have only a few repeats, using the average of the available replicate measurements for each subject may not be sufficient to adjust for the measurement error. The methods proposed in this paper can be applied to provide additional correction procedures in the analysis of case-control data where subjects have a variable number of repeated measures of the exposure. The advantage of our algorithm is that it is conceptually straightforward and relatively easy to implement, especially with the amount of computing power that is now readily available to most investigators.

A manuscript based on this work is being prepared for submission for publication. Prior to submission, however, additional work on the bootstrap algorithm will be performed. An alternative bootstrap sampling procedure, in which the same sample, generated using the matched set as the sampling unit, is utilized for estimation of both the variance components of the reliability coefficient as well as the conditional logistic regression parameter. This sampling approach may yield a more valid estimate of the corrected confidence intervals than the above approach.

Table 1: Reproducibility of Total, %Free, and %SHBG-Bound Estradiol, Adjusted for Case/Control Status and Matching Stratum

Hormone	Within-Subject Variance	Between-Subject Variance	Reliability Coefficient
Estradiol	0.17	0.15	0.47
% Free Estradiol	0.0168	0.033	0.67
% SHBG-Bound Estradiol	9.38	99.48	0.91

Table 2: Corrected and Uncorrected Logistic Regression Parameter Estimates, Confidence Intervals, and Odds Ratios for the Associations of Total, % Free, and % SHBG-bound Estradiol Level and Risk of Breast Cancer

Exposure Variable	Regression Coefficient	95% C.I.	Odds Ratio*
Total Estradiol†			
Uncorrected first measurement	0.66	(0.24 - 1.09)	3.02
Uncorrected average	0.77	(0.32 - 1.22)	3.60
Corrected average	1.37	(0.55 - 3.08)	9.70
% Free Estradiol			
Uncorrected first measurement	1.70	(0.69 - 2.71)	3.07
Uncorrected average	1.73	(0.70 - 2.77)	3.13
Corrected average	2.47	(1.19 - 4.10)	5.10
% SHBG-Bound Estradiol			
Uncorrected first measurement	-0.046	(-0.068 - -0.024)	0.25
Uncorrected average	-0.045	(-0.067 - -0.023)	0.26
Corrected average	-0.050	(-0.070 - -0.026)	0.22

* Comparing women at 90^{th} vs. 10^{th} percentile of observed distribution

† Total estradiol measurements were log-transformed

References

Armstrong, B.G. The effects of measurement errors on relative risk regressions. Am J Epidemiol 1990; 132:1176-1184.

Armstrong, B., Whittemore, A., and Howe, G. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. Statistics in Medicine 1989; 8:1151-1163.

De Klerk, N.H., English, D.R., and Armstrong, B.K. A review of the effects of random measurement error on relative risk estimates in epidemiological studies. International Journal of Epidemiology 1989; 18:705-712.

Efron, B. and Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1986; 1:54-77.

Haukka, J.K. Correction for covariate measurement error in generalized linear models–a bootstrap approach. Biometrics 1995; 51:1127-1132.

Kim, M.Y., Pasternack, B.P., Carroll, R.J., Koenig, K.L., and Toniolo, P.G. Estimating the reliability of an exposure variable in the presence of confounders. Statistics in Medicine 1995; 14:1437-1446.

Liu, X. and Liang, K.Y. Efficacy of repeated measures in regression models with measurement error. Biometrics 1992; 48:645-654.

Prentice, R.L. Covariate measurement error and parameter estimation in a failure time

regression model. Biometrika, 1982; 69:331-342.

Rosner, B., Spiegelman, D., Willett, W.C. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990; 132:734-745.

Rosner, B., Spiegelman, D., Willett, W.C. Correction of logisitic regression relative risk estimates and confidence intervals for random within-person measurement error. Am J Epidemiol 1992; 136:1400-1413.

Thomas, D., Stram, D., and Dwyer, J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. Annual Review of Public Health 1993; 14:69-93.

Toniolo, P., Pasternack, B., Shore, R., Sonnenschein, E., Koenig, K., Rosenberg, C., Strax, P., Strax, S. Endogenous hormones and breast cancer: A prospective cohort study. Breast Cancer Res Treat 1991; 18:S23-S26.

Toniolo, P., Koenig, K., Pasternack, B., Banerjee, S., Rosenberg, C., Shore, R., Strax, P., and Levitz, M. Reliability of measurements of total, protein bound, and unbound estrdiol in serum. Cancer Epidemiology, Biomarkers and Prevention (1994) 3:47-70.

Toniolo, P., Levitz, M., Zeleniuch-Jacquotte, A., Banerjee, S., Koenig, K., Shore, R., Strax, P., Pasternack, B. A prospective study of endogenous estrogens and breast cancer in post-menopausal women. Journal of the National Cancer Institute (1995) 87:190-197.

Whittemore, A.S. Errors-in-variables regression using stein estimates. The American Statis-

tician 1989; 43:4.

Chapter II

Adjusting Premenopausal Estradiol Levels for Day of Menstrual Cycle using Cubic Splines

1 Introduction

Although levels of prolactin and bioavailable estradiol appear to be relatively stable over the phases of a woman's menstrual cycle, other hormones, such as total estradiol, fluctuate considerably. (Toniolo et al., 1993; Koenig et al., 1993.; Wu et al., 1976; Takatani et al., 1991). Thus, studies investigating the association of total estradiol and risk of breast cancer among premenopausal women must adjust the hormone level for day of cycle either in the design or analysis stage of the study.

In the NYU Women's Health Study, a nested case-control study of serum hormonal levels and breast cancer, one of the criteria for matching controls with a breast cancer case among pre-menopausal women was day of menstrual cycle at the time of the first blood donation. More specifically, the first blood donation was matched on exact day of blood specimen relative to next expected onset of menses. Subsequent blood donations, however, could not be matched on day of cycle. Therefore, a method was needed to standardize hormone measurements obtained at different times during the menstrual cycle to make more valid comparisons between subjects in the same matched set.

Rosenberg et al (1994) used the first measurement from each control subject to fit a three-piece spline model to describe the change in estradiol levels over the menstrual cycle. For each subject, the estradiol measurement adjusted for day of cycle was then calculated as the number of standard deviations above or below the expected value from the calibration curve. The limitation with this approach, however, is that because only the first measurement from each subject was used to fit the calibration curve, the data are cross-sectional and the resulting curve reflects not only within-subject variation in the hormone level, but between-subject variation as well. Ideally, the estimated calibration curve should reflect only within-individual trends.

We propose an alternative method for describing the within-subject change in estradiol

levels over the menstrual cycle, based on a mixed-model analysis-of-variance model with splines, which utilizes the repeated measurement data for each subject. The estimated curve is then used to adjust each subject's hormone level for day-of-cycle in evaluating the association of estradiol level with risk of breast cancer among pre-menopausal women.

2 Methods

Let y_{ij} denote the hormone level of the i^{th} woman on the j^{th} occasion for $i = 1, ...n$, $j = 1, ...k_i$. Furthermore, let t_{ij} denote the number of days prior to next menses at which y_{ij} was measured. We assume a model of the form

$$y_{ij} = \mu + \alpha_i + S(t_{ij}) + \epsilon_{ij}$$

where μ denotes an overall mean, α_i denotes a random subject effect from a $N(0, \sigma_\alpha^2)$ distribution, $S(t_{ij})$ is cubic spline function, and the ϵ_{ij} are independent errors from a $N(0, \sigma_e^2)$ distribution. We further assume that the subject effects and the error terms are mutually independent. The presence of the random subject effect in the model accounts for the correlation between repeated measurements on the same subject.

We chose to use cubic splines to model estradiol levels versus day of cycle because this method provides great flexibility in fitting models, is visually smooth, and requires fewer constants to fit than higher degree splines. Rosenberg et al utilized two parabolic and one linear function to describe the change in estradiol over the menstrual cycle, with only a single continuity restriction. Thus, although their overall function was continuous, it was not smooth at the two join points.

When fitting a cubic spline model, more join points or knots are better if the variable changes quickly over the covariate space. However, too many knots can lead to over-fitting of the data and more parameters to estimate. Stone (1986) suggested that 5 knots should provide enough flexibility for a reasonable number of degrees of freedom.

Given that the average length of a menstrual cycle is 28 days, we positioned 5 knots at the 5 day intervals: 5, 10, 15, 20, and 25 days prior to next menses. Using the + notation of Smith (1979), let

$$u_+ = u \quad \text{if} \quad u > 0$$

$$u_+ = 0 \quad \text{if} \quad u \leq 0.$$

Then the cubic spline can be specified as:

$$\begin{aligned} S(t) &= \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 (t-5)_+^3 + \beta_5 (t-10)_+^3 \\ &\quad + \beta_6 (t-15)_+^3 + \beta_7 (t-20)_+^3 + \beta_8 (t-25)_+^3 + \epsilon. \end{aligned}$$

It follows that the overall mixed ANOVA-model has the following form:

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \beta_4 (t_{ij}-5)_+^3 + \beta_5 (t_{ij}-10)_+^3 \\ &\quad + \beta_6 (t_{ij}-15)_+^3 + \beta_7 (t_{ij}-20)_+^3 + \beta_8 (t_{ij}-25)_+^3 + \epsilon_{ij}. \end{aligned}$$

This model assumes that the shape of the function describing the change in estradiol over the menstrual cycle is the same for all subjects, but that subjects can differ with regard to the intercept term.

The mixed ANOVA-model was fit using maximum likelihood methods from the SAS PROC MIXED procedure. 498 estradiol measurements from 367 control subjects were utilized in the analysis. 278 subjects had 1 measurement, 60 had 2, 28 had 3, and 4 had 4 measurements. Only measurements obtained less than 35 days prior to next menses were included. Estradiol levels were log transformed prior to model fitting.

The estimated mean curve describing the change in log estradiol level over the menstrual cycle is shown in Figure 1. A subject's hormone level adjusted for day of cycle can then be expressed as the deviation of the subject's observed value from the expected value for that day of the cycle based on the fitted curve:

$$x_{ij} = y_{ij} - \hat{S}(t_{ij}).$$

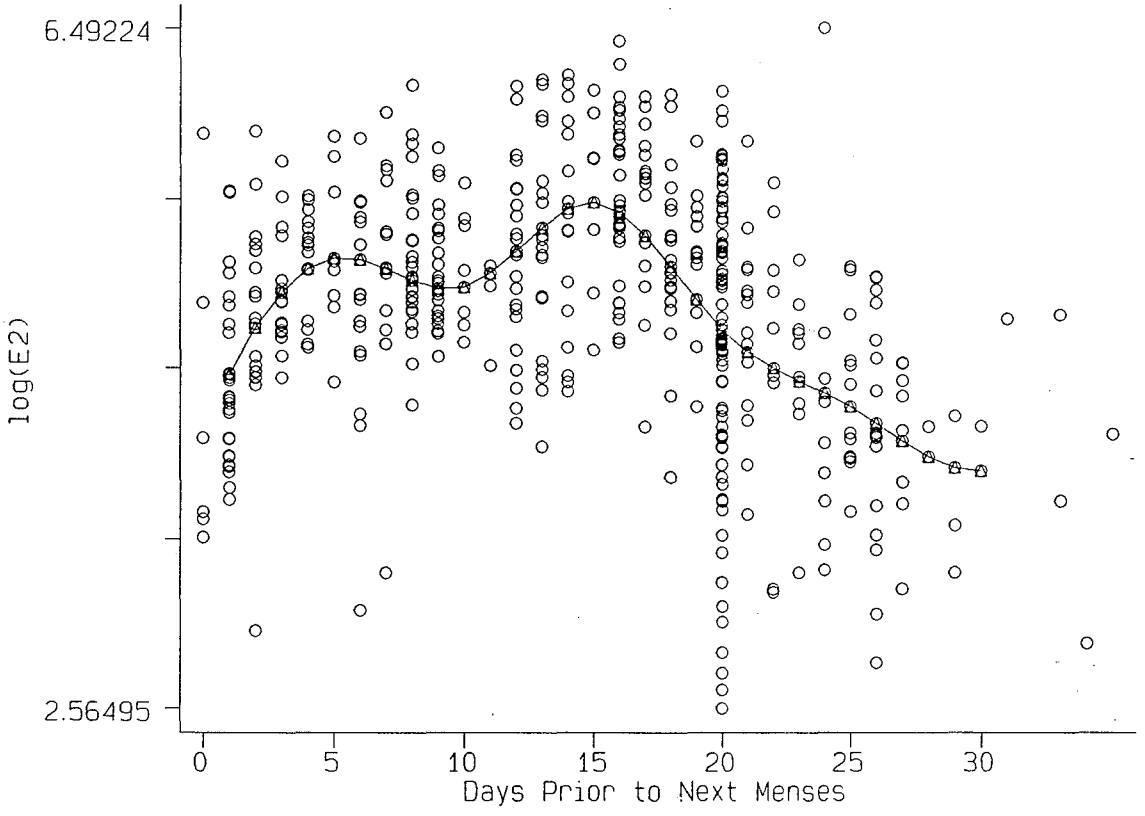
Similarly, when the average hormonal level is used as the exposure, the adjusted average can be calculated as:

$$\bar{x}_{i.} = \{\sum_j y_{ij} - \hat{S}(t_{ij})\}/n.$$

3 Conclusions

This work is still in progress. The next phase will focus on fitting conditional logistic regression models to total estradiol levels adjusted for day of cycle, using the estimated calibration curve from above. The resulting estimates of relative risk will be compared with the estimates based on the unadjusted exposure variables.

Although much of the within-subject variability in the estradiol levels adjusted for day of cycle will be reduced since the variability due to cycle will in principle be eliminated, the remaining within-subject variance, due to laboratory measurement error and short-term fluctuations, may still be non-negligible. Methods for correcting for errors-in-measurement in the adjusted hormonal levels will also be explored.


6.49224
2.56495
log(E2)
0
5
10
15
20
25
30
Days Prior to Next Menses

References

Koenig, K.L., Toniolo, P.G., Bonfrer, P.F., et al. Reliability of serum prolactin measurements. Cancer Epidemiology, Biomarkers and Prevention (1993) 137:1068-1080.

Rosenberg, C.R., Pasternack, B.S., Shore, R.E., et al. Premenopausal estradiol levels and the risk of breast cancer: a new method of controlling for day of the menstrual cycle. Am J Epidemiol (1994) 140: 518-525.

Smith, P.L. Splines as a useful and convenient tool. The American Statistician (1979) 33: 57-62.

Stone, C.J. Comment on Hastie and Tibshirani. Statistical Science (1986) 1: 312-34.

Takatani, O., Okumoto, T., Kosano, H. Genesis of breast cancer in Japanese: A possible relationship between sex hormone binding globulin (SHBG) and serum lipid components. Breast Cancer Res Treat (1991) 18:s27-s29.

Toniolo, P., Koenig, K., Pasternack, B., et al. Reliability of measurements of total, protein bound, and unbound estrdiol in serum. Cancer Epidemiology, Biomarkers and Prevention (1994) 3:47-70.

Wu, C.H. Free and protein-bound plasma estradiol-17B during the menstrual cycle. J Clin Endocrinol Metab (1976) 43:436-445.

Chapter III

Sample Size and Study Design Considerations for Half-Life Studies

1 Introduction

The accumulation of PCBs (polychlorinated biphenyls) and DDE (1,1 dichloro-2,2-bis(p-chlorophenyl) ethylene) residues, and other environmental contaminants in the body may potentially have adverse health effects. Individuals who are able to clear these toxic compounds from the body at a faster rate, and thus have shorter half-lives, may be at lower risk of diseases associated with the toxins. Thus, in order to fully elucidate the role of environmental contaminants in the development of disease, their rates of persistence in the body must be accurately quantified.

Previous studies estimating the half-life of PCBs have yielded inconsistent results, however. Reported estimates of half-life range from .5 months to 17 years for PCB mixtures. (Yakushiji et al., 1984; Phillips et al., 1989; Elo et al 1985; Lawton et al., 1985). For specific PCB components, half-lives have been estimated to be from less than 1 year to about 30 years (Yakushiji et al, 1984; Chen et al, 1982). Similarly, data on the half-life of DDE are variable and limited.

The lack of consistency among study estimates of half-life may be largely due to the small sample sizes and limited number of repeated measurements per subject utilized in these studies. For example, Chen et al (1982) examined the rates of elimination of PCBs from the blood of PCB-poisoned subjects in Taiwan using two to three serial blood samples from 17 individuals taken over a period of 6-14 months. Similarly, Steele et al (1986) calculated the half-life of PCBs using two measurements of PCB concentrations made 7 years apart.

Phillips (1989) investigated how analytical (laboratory) error and the time interval between measurements affect the variability and possible bias in estimates of half-life calculated from two measurements. Results indicate that half-life estimates based on only two measurements become increasingly variable at shorter time intervals between measurements and at higher levels of analytical error.

The precision of half-life estimates, however, is not only dependent on the magnitude of

analytical error and the time interval between measurements, but also on the number of repeated measurements utilized in the estimation procedure. Given laboratory cost contraints, time constraints, and other limitations on the physical resources of a study on half-life, investigators must decide where to allocate the resources in order to obtain the most precise estimate of half-life.

Issues of sample size and study design for estimating subject-specific, as well as population half-lives of environmental contaminants have not been formally addressed in the environmental and epidemiologic literature. The objectives of this paper are to provide useful guidelines for choosing the number of repeats and the optimal time interval between repeats needed for estimating an individual's half-life with a given level of precision, while minimizing the cost of the study. In addition, sample size and power considerations for studies comparing the population half-lives between two groups will be investigated. An example is presented using data from a study on PCBs and breast cancer.

2 Methods

For most environmental toxins, the rate of elimination from the body may be described by the following one-compartment exponential decay model:

$$C(t) = C_0 e^{-\lambda t}, \tag{1}$$

where C(t) is the concentration of the toxin at time t, C_0 denotes the initial concentration, and λ is the rate constant. The half-life, $t_{1/2}$, which is the time after which the level of toxin is reduced to half its original value, is equal to $\ln(2)/\lambda$.

If both sides of (1) are log-transformed, then we have the linear relationship:

$$\ln\{C(t)\} = \ln(C_0) - \lambda t. \tag{2}$$

Thus, given $\mathbf{C(t)} = \{C(t_1), ..., C(t_k)\}$, the set of serial measurements of the toxin obtained on a subject at times, $\{t_1, ..., t_k\}$, the rate constant, λ, may be estimated from the slope

of the linear regression of $\ln\{\mathbf{C(t)}\}$ versus $\mathbf{t}$. The least-squares estimate of λ is equal to $\hat{\lambda} = \sum_{j=1}^{k}\{\ln\{C(t_j)\} - \bar{C}\}(t_j - \bar{t})/\sum_{j=1}^{k}(t_j - \bar{t})^2$, where $\bar{C}$ and $\bar{t}$ denote the average logarithm of the level of toxin and average time of measurement, respectively. The corresponding half-life for the subject may be estimated by $\hat{t}_{1/2} = \ln(2)/\hat{\lambda}$.

The sample size and study design issues associated with estimating the half-life will depend on whether the focus is on obtaining a precise estimate of an individual's half-life or a population half-life. The former would be of interest, for example, in studies exploring the relationship between an individual's rate of elimination of the toxin with a particular genetic characteristic. On the other hand, a precise estimate of a population half-life would be pertinent when the investigator is interested in comparing the average half-lives between two or more groups, such as diseased and non-diseased subjects.

Study Design for Estimating Individual Half-Lives

If the goal is to estimate individual half-lives with a certain level of precision, then clearly, the number of subjects to include in the study is not relevant. The frequency of measurement and duration of follow-up are the primary factors which will determine the precision of the individual's half-life estimate. This can be shown as follows.

The variance of $\hat{\lambda}$, the least-squares estimate of the rate parameter, is equal to $\sigma_e^2/\sum_{j=1}^{k}(t_j - \bar{t})^2$, where σ_e^2 is the variance of the deviation of the observed $\ln\{\mathbf{C(t)}\}$ from the value predicted by the regression line in (2). Then, using the Delta method (Cox and Hinkley, 1974), the variance of $\hat{t}_{1/2}$ is equal to,

$$V(\hat{t}_{1/2}) = \ln(2)^2 \left(\frac{1}{\lambda}\right)^4 \frac{\sigma_e^2}{\sum_{j=1}^{k}(t_j - \bar{t})^2}. \qquad (3)$$

Let $\mathbf{t} = \{t_1, t_2, ..., t_k\}$ denote k equally spaced points in time, where the time interval between points is equal to I. Then the study duration, D, is equal to $I(k-1)$. Following the arguments in Schlesselman (1973), we can express $\sum_{j=1}^{k}(t_j - \bar{t})^2$ as a function of D and

k:

$$\sum_{j=1}^{k}(t_j - \bar{t})^2 = \frac{D^2 k(k+1)}{\{12(k-1)\}}.$$

It follows that the variance of $\hat{t}_{1/2}$ can be expressed as:

$$V(\hat{t}_{1/2}) = \ln(2)^2\left(\frac{1}{\lambda}\right)^4 \frac{\sigma_e^2\{12(k-1)\}}{D^2 k(k+1)} = \left(\frac{t_{1/2}}{\lambda D}\right)^2 \frac{\sigma_e^2\{12(k-1)\}}{k(k+1)}. \quad (4)$$

Thus, (4) describes how the precision of $\hat{t}_{1/2}$ is a function of the study duration, D, the number of repeated measurements on a subject, k, σ_e, and λ. For fixed values of the underlying rate parameter, λ, and σ_e, the variance of $\hat{t}_{1/2}$ is directly proportional to $\omega = \frac{ln(2)^2 12(k-1)}{D^2 k(k+1)}$. Schlesselman presented tables which show how the precision of a slope changes over different values of k and D. Table 1 describes analogous results for the precision of the half-life. Specifically, we calculated ω for various values of k and D. One can easily see how ω, and thus, the variance of the half-life, decreases as the number of repeats and the duration of study increases. The exception, however, is that for a fixed duration of study, obtaining 3 measurements does not result in additional precision compared with 2 measurements. (This is due to the algebraic result that the term $(k-1)/k(k+1)$ in (4) is the same for $k = 2$ or 3.) Furthermore, for large k, the variance of $\hat{t}^i_{1/2}$ is proportional to $1/(D^2 k)$. Thus, a unit increase in the duration of the study will result in greater precision of the half-life estimate than a unit increase in the number of repeated measurements. Finally, note that some combinations of k and D will yield the same level of precision. For example, 10 measurements obtained over 7 months result in the same precision as 7 measurements over 8 months, and 3 measurements over 10 months.

The choice between different pairs of (k, D) for estimating the half-life will depend upon the relative costs of each measurement and each time interval of follow-up (which may include staff salaries and other administrative costs). If the two costs are equivalent, then results from Table 1 suggest that resources should be directed toward extending the duration of the study, since this will result in larger gains in precision than will increasing the number of

measurements. When the costs of (k, D) differ, however, the allocation of resources which will result in the most precise estimate of $t_{1/2}$ is less clear.

For each subject, let $C = c_1 k + c_2 D$ equal the total cost of measuring the subject k times over a duration of D years, where c_1 denotes the cost of an individual measurement, and c_2 denotes the cost per year of follow-up. Assume that the goal is to estimate an individual's half-life with variance equal to V, while minimizing the total cost per study subject. If we make the simplifying assumption that for large k,

$$V \approx \ln(2)^2 \left(\frac{1}{\lambda}\right)^4 \frac{12\sigma_e^2}{D^2 k} = \left(\frac{t_{1/2}}{\lambda D}\right)^2 \frac{12\sigma_e^2}{k}, \tag{5}$$

then a Lagrange multiplier may be used to minimize C subject to the constraint in (5). After some algebraic manipulations, we have the result

$$k = \left\{\frac{\ln(2)^2 3 c_2^2 \sigma_e^2}{V c_1^2 \lambda^4}\right\}^{1/3} = \left\{\frac{3\sigma_e^2}{V}\left(\frac{c_2}{c_1}\right)^2 \left(\frac{t_{1/2}}{\lambda}\right)^2\right\}^{1/3}, \tag{6}$$

and

$$D = \left\{\frac{\ln(2)^2 24 c_1 \sigma_e^2}{V c_2 \lambda^4}\right\}^{1/3} = \left\{\frac{24\sigma_e^2}{V}\left(\frac{c_1}{c_2}\right) \left(\frac{t_{1/2}}{\lambda}\right)^2\right\}^{1/3}, \tag{7}$$

as the optimal values of k and D which will minimize the cost for a specified level of precision, V. As expected, the optimal k and D depend on c_2/c_1, the ratio of the cost per month of follow-up to the cost per measurement. As this ratio increases, the optimal design favors increasing the number of repeated measurements and decreasing the duration of follow-up. In order to calculate k and D from (6) and (7), respectively, values of λ and σ_e^2 must be assumed. Estimates may be obtained from the literature or preliminary studies.

The above result is valid only when k is large enough so that $(k-1)/(k+1) \approx 1$. When this assumption does not hold, closed form solutions are not available for calculating the optimal k and D, and iterative methods must be utilized. Investigators who are unfamiliar with iterative numerical techniques may need to consult a statistician.

Study Design for Estimating and Comparing Population Half-Lives

In the above discussion, it was assumed that the primary focus was on estimating the subject-specific half-lives. Thus, the size of the study population was not relevant. However, when the goal is to estimate the average half-life in a particular population, or to compare the half-lives in two different populations, then one needs to consider the number of subjects to include in the study, in addition to the frequency and duration of measurements.

Assume that the sample population is comprised of N subjects, and that each subject has a "true" rate parameter, λ_i, which is distributed with mean λ_P and variance, σ_s^2. Thus, λ_P can be interpreted as the underlying population rate parameter, and σ_s^2 is the variance in λ_i between individuals. Furthermore, assume that the frequency of measurement, study duration, and σ_e are the same for all subjects.

Given the estimated subject-specific half-lives: $\{\hat{t}^1_{1/2}, ..., \hat{t}^N_{1/2}\}$, the population half-life, $t^P_{1/2}$, may be estimated by: $\hat{t}^P_{1/2} = \{\hat{t}^1_{1/2} + ... + \hat{t}^N_{1/2}\}/N$. Using result (4) and the assumptions above, it can be shown that the variance of $\hat{t}^P_{1/2}$ is equal to

$$V(\hat{t}^P_{1/2}) = \ln(2)^2(\frac{1}{\lambda_P})^4 \left[\sigma_s^2 + \frac{\sigma_e^2\{12(k-1)\}}{D^2k(k+1)}\right]\frac{1}{N}. \qquad (8)$$

Equation (8) can be used to determine the k, D, and N which will result in a certain level of precision in the population half-life estimate. One can see from the form of the equation that the precision of $\hat{t}^P_{1/2}$ improves as k, D, and N increase, and that increases in N will diminish both the contributions of σ_s^2 and σ_e^2 to the variance. Note that the variance is no longer directly proportional to a factor which is a function only of k, D and N. Thus, tables similar to Table 1 cannot be generated unless values for σ_e^2 and σ_s^2 are assumed. The use of (8) will be illustrated in the example.

Design issues for studies comparing the half-lives between two populations will now be considered. Let $t^1_{1/2}$ and $t^2_{1/2}$ denote the half-lives in the two populations. The null hypothesis is $H_0 : t^1_{1/2} = t^2_{1/2}$. We assume that the sample sizes in both groups are equal to N, that all subjects have the same number of repeated measurements obtained at the same time intervals, and that the between-subject variance of the true rate parameter is equal to σ_s^2 for

both populations. It is shown in Appendix I that for fixed values of k and D, the required number of subjects per group for attaining a $(1-\beta)$ level of power to detect the alternative hypothesis, $H_A : t^1_{1/2} \neq t^2_{1/2}$ at an α significance level is

$$N = \ln(2)^2 \left(\frac{z_{\alpha/2}\sqrt{2/\bar{\Lambda}^4} + z_\beta \sqrt{\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4}}}{(t^1_{1/2} - t^2_{1/2})} \right)^2 \left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)} \right], \tag{9}$$

where $z_{\alpha/2}$ and z_β denote the standard normal deviates corresponding to $\alpha/2$ and β significance levels, respectively, and $\bar{\Lambda} = (\lambda_1 + \lambda_2)/2$.

Note that since the required sample size depends on λ_1 and λ_2, the actual values of $t^1_{1/2}$ and $t^2_{1/2}$ need to be specified, and not just the magnitude of their difference. Equation (9) can also be easily re-expressed to determine the k or D to attain a specified level of power, for fixed values of the other parameters.

The formula for determining the sample size was derived assuming that the duration of the study and the number of repeats are fixed. However, the most common situation when designing a study is that k and D, in addition to N, need to be determined. Methods similar to the above may be utilized to calculate the optimal values for the number of subjects, number of repeats, and duration of study which will minimize the overall study cost, while attaining a specified level of power. The total cost of the study can be denoted as $C = c_0 + (c_1 k + c_2 D + c_3)2N$, where c_0 denotes overhead and other fixed costs which are independent of k, D, and N; c_1 and c_2 are the costs associated with each measurement and each interval of follow-up, respectively; and c_3 denotes the cost of enrolling each additional subject.

The optimal parameter values for k, D, and N can be determined by minimizing C, subject to the constraint in (9). Unlike the previous problem, however, this has no closed form solution and must be solved iteratively. A Newton-Raphson algorithm, written in SAS PROC IML, was utilized to estimate the optimal parameters (Press et al., 1986). This algorithm requires calculation of the first and second order derivatives, with respect to the

parameters of interest, of the function which is to be minimized. In this case, the function is $C = c_0 + (c_1 k + c_2 D + c_3)2N$, with N substituted by the expression in (9). Expressions for the first and second-order derivatives of C with respect to k and D, and details of the algorithm are given in Appendix II.

Specific values of $z_\alpha, z_\beta, \sigma_s^2, \sigma_e^2, t_{1/2}^1$, and $t_{1/2}^2$, as well as the costs, c_1, c_2, and c_3, must be assumed. Note that because the first and second-order derivatives of C with respect to k and D are independent of c_0, the overhead cost will not affect the outcome of the minimization process, and hence, need not be specified. Given initial starting values for k and D, the algorithm iteratively finds the values which minimize C. The optimal number of subjects, N, is then calculated from (9). An example illustrating the methods is presented in the next section.

3 Example

In this section, utilization of the methods to design a study to compare the differences in the half-life of PCBs between subjects with and without breast cancer will be illustrated. First, values of the variance components, σ_s^2, the between-subject variance in the true rate parameter, and σ_e^2, the variance of the deviations of the observed measurements (log transformed) from the values predicted from equation (2), must be assumed. Variance estimates were obtained using pilot data from the NYU Women's Health Study (NYUWHS), a prospective cohort of 14,291 women who have been donating multiple blood samples over time (Toniolo et al, 1991). A breast cancer case-control study nested in this cohort found elevated, but non-significant, levels of PCBs measured at enrollment among cases relative to controls (Wolff et al, 1993). No half-lives were measured at that time because only one blood donation per subject was analyzed. Subsequently, pilot data became available on subjects in the NYUWHS who had at least 3 blood donations. Concentrations of PCBs were measured in serum specimens that have been collected and stored since enrollment; the assays were

performed under the direction of Dr. Mary Wolff at Mt. Sinai Medical Center. Details of the experimental protocol are provided in Wolff et al (1991).

In calculating the half-lives for this cohort, the concentrations of PCBs within subjects are assumed to be decreasing over time. In principle, however, the body burden of PCBs may actually increase in individuals who are chronically exposed to low levels of the toxin and whose initial concentrations were in the range of normal background levels, resulting in negative half-life estimates. For our example, the analysis was restricted to include only the 15 subjects with at least 3 measurements of PCBs available who had a positive estimate of half-life. The mean half-life of PCBs among these subjects was estimated to be 10 years.

An estimate of σ_e^2 was obtained by fitting the following linear mixed ANOVA model:

$$Y_{ij} = \mu + \alpha_i + \lambda_i t_{ij} + e_{ij}, \tag{11}$$

where Y_{ij} is defined as the logarithm of the j^{th} measurement of PCB from subject i, μ denotes the overall mean, α_i denotes a random subject effect, λ_i is the rate parameter for subject i, t_{ij} is the time since enrollment for subject i and donation j, and e_{ij} is the residual error, which is assumed to be distributed with mean 0, and common variance, σ_e^2. The mean squared error resulting from model (11) estimates σ_e^2. Fitting (11) to the NYUWHS data yielded $\hat{\sigma}_e^2 = .046$.

Obtaining an estimate of the between-subject variance of the true rate parameters, σ_s^2 was more problematic. If the measurements from all subjects were made at the same set of time points, $\mathbf{t} = \{t_1, ..., t_k\}$, then one could estimate σ_s^2 by first estimating λ_i for all subjects and subtracting $\frac{\hat{\sigma}_e^2}{\sum_{j=1}^{k}(t_j - \bar{t})^2}$ from the observed variance of $\hat{\lambda}_i$, since the unconditional variance of $\hat{\lambda}_i$ is equal to $\sigma_s^2 + \frac{\hat{\sigma}_e^2}{\sum_{j=1}^{k}(t_j - \bar{t})^2}$. However, in the NYUWHS and in most other studies, subjects have different numbers of repeated measurements obtained at varying time intervals. In this case, a conservative estimate of σ_s^2 would be to use the observed variance of $\hat{\lambda}_i$. Although this leads to an overestimate of the required sample size, the approximation improves as the number of repeated measurements and the duration between measurements become large.

The observed variance of the rate parameters of PCBs from our pilot data was estimated to be .0028.

Before determining the optimal design for comparing the half-lives between two populations, we illustrate how one can generate tables using (8) and the estimates of σ_s^2 and σ_e^2 to evaluate the effect of increasing k, D and N on the precision of the estimate of a single population half-life. Suppose one assumes that the true underlying half-life of PCB for the breast cancer cases is 11 years. This corresponds to a population rate parameter of $\lambda_p = \ln(2)/11 = .063$. Using (8), we generated Table 2, which shows the variance of $\hat{t}_{1/2}^P$ for selected values of k, D and N. For example, with a sample size of 75 subjects measured four times over a period of 8 years, the variance of the estimated half-life will be 1.66, corresponding to a 95% confidence interval width of: $2 \times 1.96 \times \sqrt{1.66} = 5.05$ years for the true population half-life. In this particular example, increasing the duration of study by a given number of years, say x, results in greater gains in precision compared with increasing by x the number of repeats or number of subjects. This result, however, may not apply for different values of σ_s^2 and σ_e^2.

The optimal design for comparing the population half-lives of PCB between breast cancer cases and controls will now be determined. The following values for the costs of the study were assumed: \$200 for each PCB assay (c_1), \$25 for each year of follow-up (c_2), and \$75 to enroll each subject (c_3). Assuming that the half-life of PCB among control subjects is 8 years and that the study should have 80% power to detect an increase in the half-life to 11 years among breast cancer cases at an $\alpha = .05$ significance level, we found, using the iterative algorithm described in Appendix II, that the optimal design is to enroll 100 subjects per group, and to obtain 2 measurements per subject over 12 years.

Even though this design is the one which will minimize the overall cost of the study, in practice, it may not be feasible to conduct the study over a time period as long as 12 years. Suppose that 5 years is the maximum feasible duration of study. Then, one can minimize C with respect to k and N, while keeping D fixed at 5 years, to obtain the optimal design

for a 5 year study. Iterative methods similar to the above were used to determine that the optimal design for a 5 year study is to obtain 2 measurements per subject on 186 subjects per group. Although this design will yield the same level of power over a shorter duration as the first design, it will cost an additional $6275.

Figure 1 shows how the optimal k, D and N change as a function of the cost of the assay, assuming the values of the other parameters have not changed. For example, if the cost of the PCB assay were only $2 rather than $200, then the optimal design is to obtain 26 measurements per subject over 5 years and enroll 103 subjects per group. The greatest changes in the optimal values for k, D and N occur when $c1$ ranges from $1-$9. For assay costs greater than $9, the optimal value for k remains stable at 2 measurements. Corresponding changes in the optimal D and N in this region of $c1$ are minimal. Similar graphs can be generated to evaluate the impact of varying the values of the other parameters on the optimal values.

It is straightforward to show that specification of the level of power, type I error rate, and population half-lives only influence the determination of the optimal N, and not k and D (see Appendix II). Thus, in order to evaluate how the optimal design changes as a function of α, $1-\beta$, $t^1_{1/2}$, and $t^2_{1/2}$, one need only to re-calculate N using (9), since the required k and D will remain unchanged. For instance, continuing the initial example from above, in order for the study to attain 70%, as opposed to 80% power, the required number of subjects is reduced to 77 per group, while the optimal k and D remain as above ($k = 2; D = 12$). The values for k and D are affected only by the costs, c_1, c_2 and c_3, and the values of the variance components, σ^2_s and σ^2_e.

4 Conclusions

Understanding the pharmacokinetics, and in particular, the rate of excretion from the body of environmental contaminants is crucial for ascertaining the etiologic role of these risk factors

in the development of disease. In this paper, methods for designing studies on estimating and comparing the half-lives of environmental toxins have been described. The ability to utilize these methods, however, may be limited by the availability of preliminary estimates for the variance components. Although most studies on population half-lives provide estimates of the variance of the population rate parameters, which may be used as an upper bound estimate of σ_s^2, estimates of σ_e^2 are rarely published. The availability of pilot data becomes especially important in this case. Also, because iterative methods are required to determine the optimal design for comparing two population half-lives, the techniques may not be easily implemented in practice for some investigators and a statisician may need to be consulted. Finally, the techniques in this paper are based on the assumptions of a one-compartment exponential decay model and a linear least-squares regression estimate of the rate parameter, λ. Thus, they cannot be applied to the multi-compartment case. Extension of this work to accommodate the multi-compartment assumption will be the subject of future research.

Most published reports on the half-lives of environmental contaminants have been based on small numbers of subjects and small numbers of repeated measurements. The large variability in the published estimates of the half-lives of toxins such as PCB may reflect the lack of precision that results from inadequate study designs. This paper demonstrates the gains in precision and statistical power that may be achieved by increasing the sample size, number of repeats, and time interval between repeats, and underscores the importance of study design when planning studies on half-life.

Table 1: Values of ω ($\times 10$) as a function of the number of repeats, k, and the duration of study, D.

	D									
k	1	2	3	4	5	6	7	8	9	10
2	9.609	2.402	1.068	0.601	0.384	0.267	0.196	0.150	0.119	0.096
3	9.609	2.402	1.068	0.601	0.384	0.267	0.196	0.150	0.119	0.096
4	8.648	2.162	0.961	0.541	0.346	0.240	0.176	0.135	0.107	0.086
5	7.687	1.922	0.854	0.480	0.307	0.214	0.157	0.120	0.095	0.077
6	6.864	1.716	0.763	0.429	0.275	0.191	0.140	0.107	0.085	0.069
7	6.177	1.544	0.686	0.386	0.247	0.172	0.126	0.097	0.076	0.062
8	5.605	1.401	0.623	0.350	0.224	0.156	0.114	0.088	0.069	0.056
9	5.125	1.281	0.569	0.320	0.205	0.142	0.105	0.080	0.063	0.051
10	4.717	1.179	0.524	0.295	0.189	0.131	0.096	0.074	0.058	0.047
15	3.363	0.841	0.374	0.210	0.135	0.093	0.069	0.053	0.042	0.034
20	2.608	0.652	0.290	0.163	0.104	0.072	0.053	0.041	0.032	0.026

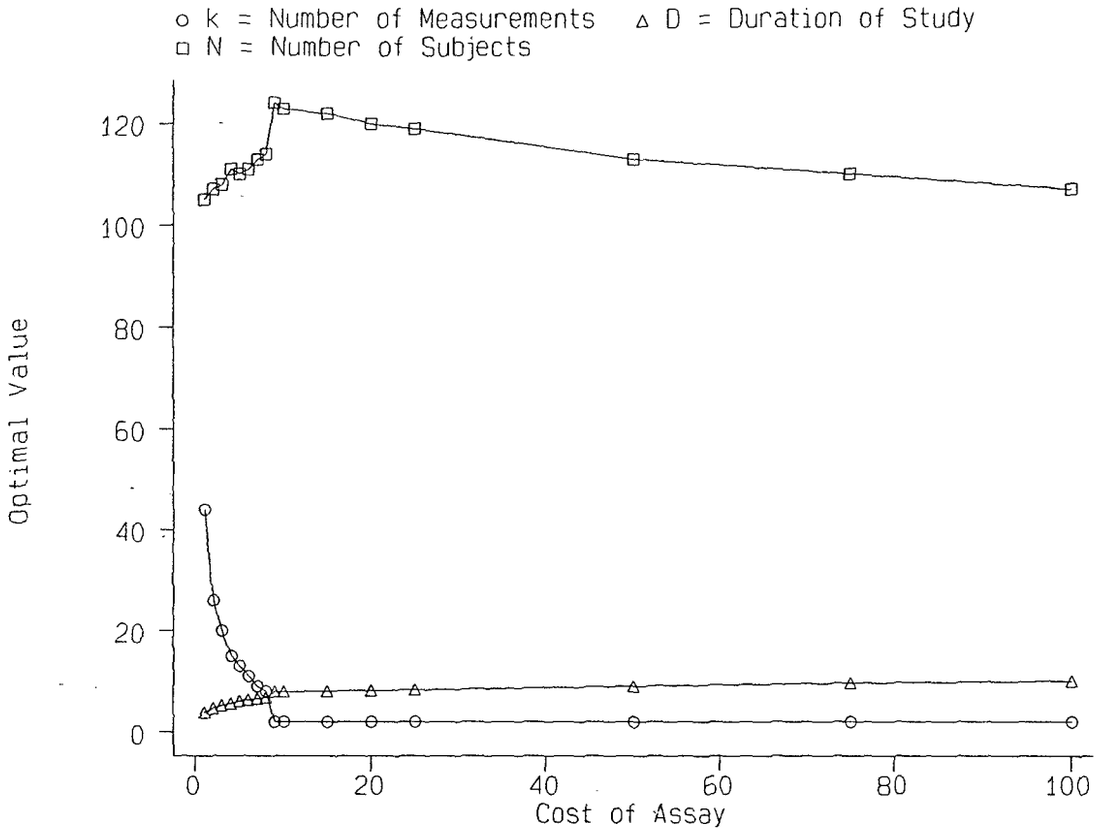
Table 2: Values of $V(\hat{t}^P_{1/2})$ for $N = 25, 50, 75, 100$; $\sigma^2_s = .0028$; $\sigma^2_e = .046$; $\lambda_p = .063$, as a function of the number of repeats, k, and duration of study, D.

N = 25							
	D						
k	2	4	6	8	10	15	20
2	31.48	10.43	6.53	5.17	4.54	3.91	3.70
4	28.67	9.73	6.22	4.99	4.43	3.86	3.67
6	23.46	8.43	5.64	4.67	4.22	3.77	3.62
8	19.78	7.51	5.23	4.44	4.07	3.71	3.58
10	17.19	6.86	4.95	4.28	3.97	3.66	3.55
15	13.24	5.87	4.51	4.03	3.81	3.59	3.51
20	11.03	5.32	4.26	3.89	3.72	3.55	3.49

N = 50							
	D						
k	2	4	6	8	10	15	20
2	15.74	5.22	3.27	2.58	2.27	1.96	1.85
4	14.33	4.86	3.11	2.50	2.21	1.93	1.83
6	11.73	4.21	2.82	2.33	2.11	1.89	1.81
8	9.89	3.75	2.62	2.22	2.04	1.85	1.79
10	8.60	3.43	2.47	2.14	1.98	1.83	1.78
15	6.62	2.94	2.25	2.01	1.90	1.80	1.76
20	5.52	2.66	2.13	1.95	1.86	1.78	1.75

N = 75							
	D						
k	2	4	6	8	10	15	20
2	10.49	3.48	2.18	1.72	1.51	1.30	1.23
4	9.56	3.24	2.07	1.66	1.48	1.29	1.22
6	7.82	2.81	1.88	1.56	1.41	1.26	1.21
8	6.59	2.50	1.74	1.48	1.36	1.24	1.19
10	5.73	2.29	1.65	1.43	1.32	1.22	1.18
15	4.41	1.96	1.50	1.34	1.27	1.20	1.17
20	3.68	1.77	1.42	1.30	1.24	1.18	1.16

N = 100							
	D						
k	2	4	6	8	10	15	20
2	7.87	2.61	1.63	1.29	1.13	0.98	0.92
4	7.17	2.43	1.56	1.25	1.11	0.97	0.92
6	5.86	2.11	1.41	1.17	1.05	0.94	0.90
8	4.95	1.88	1.31	1.11	1.02	0.93	0.89
10	4.30	1.71	1.24	1.07	0.99	0.92	0.89
15	3.31	1.47	1.13	1.01	0.95	0.90	0.88
20	2.76	1.33	1.07	0.97	0.93	0.89	0.87


o k = Number of Measurements
△ D = Duration of Study
□ N = Number of Subjects
Optimal Value
120
100
80
60
40
20
0
0
20
40
60
80
100
Cost of Assay

References

Chen, P.H., Luo, M.L., Wong, C.K., Chen, C.J. (1982) Comparative rates of elimination of some individual polychlorinated biphenyls from the blood of PCB-poisoned patients in Taiwan. Fd Chem Toxicol 20:417-425.

Cox, D.R. and Hinkley, D.V. (1974) Theoretical Statistics. New York: Chapman and Hall.

Elo O., Vuojolahti, P., Janhunen, H., Rantanen, J. (1985) Recent PCB accidents in Finland. Environ Health Persp 60:315-319.

Lawton, R.W., Brown, J.F. Jr., Ross, M.R., Feingold, J. (1985) Comparability and precision of serum PCB measurements. Arch Environ Health 40:29-37.

Phillips, D.L., Smith, A.B., Burse, V.W., Steele, G.K., Needham, L.L., Hannon, W.H. (1989) Half-life of polychlorinated biphenyls in occupationally exposed workers. Arch Environ Health 44:351-354.

Phillips, D.L. (1989) Propagation of error and bias in half-life estimates based on two measurements. Arch Environ Contam Toxicol 18:508-514.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T. Numerical Recipes. Cambridge University Press, 1986, Cambridge.

Schlesselman, J.J. (1973) Planning a longitudinal study: frequency of measurement and study duration. J Chron Dis 26:561-570.

Steele, G., Stehr-Green, P., Welty, E. (1986) Estimates of the biologic half-life of polychlorinated biphenyls in human serum. New Engl J Med 314:926-927.

Toniolo, P., Pasternack, B., Shore, R., Sonnenschein, E., Koenig, K., Rosenberg, C., Strax, P., Strax, S. Endogenous hormones and breast cancer: a prospective cohort study. Breast Cancer Res Treat 1991; 18:S23-S26.

Wolff, M.S., Rivera, M., Baker, D.B. (1991) Detection limits of organochlorine pesticides and related compounds in blood serum. Bull Environ Contam Toxicol 47:499-503.

Wolff, M.S., Toniolo, P.G., Lee, E.W., and Dubin, N. (1993) Blood levels of organochlorine residues and risk of breast cancer. J Natl Cancer Inst 85:648-652.

Yakushiji, T., Watanabe I., Kuwabara, K., Tanaka, R., Kashimoto, T., Kunita, N., Hara, I. (1984) Rate of decrease and half-life of polychlorinated biphenyls in the blood of mothers and their children occupationally exposed to PCBs. Arch Environ Contam Toxicol 13:341-345.

Appendix I: Determination of the Sample Size for Comparing Two Population Half-lives

We assume that the sample sizes from the two populations are the same and are equal to N. Let $\hat{t}^1_{1/2}$ and $\hat{t}^2_{1/2}$ denote the observed half-lives in the two populations, and $\hat{\lambda}_1$ and $\hat{\lambda}_2$ denote the estimates of the corresponding rate parameters. Then the test statistic for evaluating $H_0 : t^1_{1/2} = t^2_{1/2}$ is of the form:

$$Z = \frac{\hat{t}^1_{1/2} - \hat{t}^2_{1/2}}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (2/\bar{\lambda}^4)}},$$

where Z is distributed as $N(0,1)$, and $\bar{\lambda} = (\hat{\lambda}_1 + \hat{\lambda}_2)/2$.

If the test statistic is to have power $(1-\beta)$ to detect the alternative hypothesis, $H_A : t^1_{1/2} > t^2_{1/2}$ at a 1-sided $\alpha = .05$, then we have the following expression:

$$Pr\left\{ \frac{\hat{t}^1_{1/2} - \hat{t}^2_{1/2}}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (2/\bar{\lambda}^4)}} > z_\alpha \middle| H_A \right\} = 1 - \beta, \qquad (1)$$

where z_α denotes the critical value corresponding to the α proportion in the upper tail of the standard normal distribution.

After some algebra, (1) can be re-expressed as:

$$\begin{aligned} 1-\beta &= Pr\left\{ \frac{\hat{t}^1_{1/2} - \hat{t}^2_{1/2} - (t^1_{1/2} - t^2_{1/2})}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4})}} \right. \\ &> \left. \frac{z_\alpha \sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (2/\bar{\lambda}^4)} - (t^1_{1/2} - t^2_{1/2})}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4})}} \right\} \end{aligned}$$

Under the alternative hypothesis, the expression on the left-hand side of the inequality has a $N(0,1)$ distribution. Thus,

$$1 - \beta = Pr\left\{ Z > \frac{z_\alpha \sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (2/\bar{\lambda}^4)} - (t^1_{1/2} - t^2_{1/2})}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4})}} \right\}.$$

Note that the definition of $\bar{\lambda}$ requires knowledge of $\hat{\lambda}_1$ and $\hat{\lambda}_2$, which are available only after completion of the study. However, for large n, $\bar{\lambda}$ may be well approximated by $\bar{\Lambda} = (\lambda_1 + \lambda_2)/2$. After substituting $\bar{\Lambda}$ for $\bar{\lambda}$ above, setting the expression on the right-hand side equal to $-z_\beta$ and solving for N, we have

$$N = \ln(2)^2 \left[\frac{z_\alpha \sqrt{2/\bar{\Lambda}^4} + z_\beta \sqrt{\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4}}}{(t^1_{1/2} - t^2_{1/2})} \right]^2 \left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)} \right]. \tag{2}$$

This sample size was derived under the assumption of a one-sided alternative hypothesis. When H_A is two-sided, the required sample size is obtained by simply substituting $z_{\alpha/2}$ for z_α in (2).

Appendix II: Details of the Newton-Raphson Algorithm

The overall cost of the study is equal to :

$$\begin{aligned} C &= c_0 + (c_1 k + c_2 D + c_3)2N \\ &= c_0 + (c_1 k + c_2 D + c_3)2\left\{\ln(2)^2 \left[\frac{z_\alpha\sqrt{2/\bar{\Lambda}^4} + z_\beta\sqrt{\frac{1}{\lambda_1^4}+\frac{1}{\lambda_2^4}}}{(t_{1/2}^1 - t_{1/2}^2)}\right]^2 \left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right]\right\}, \end{aligned}$$

with N substituted by the expression in (10). The optimal k and D which will minimize C are the values which will solve the following first derivative equations:

$$\frac{\partial C}{\partial k} = A\left\{c_1\sigma_s^2 + \frac{c_1\sigma_e^2 12}{D^2}\left[\frac{2}{(k+1)^2}\right] + \left[\frac{c_3\sigma_e^2 12}{D^2} + \frac{c_2\sigma_e^2 12}{D}\right]\left[\frac{1+2k-k^2}{(k^2+k)^2}\right]\right\} = 0$$

$$\frac{\partial C}{\partial D} = A\left\{c_2\sigma_s^2 - \frac{12\sigma_e^2(k-1)}{k(k+1)}\left(\frac{2c_1 k}{D^3} + \frac{c_2}{D^2} + \frac{2c_3}{D^3}\right)\right\} = 0,$$

where $A = 2\ln(2)^2\left[\frac{z_\alpha\sqrt{2/\bar{\Lambda}^4}+z_\beta\sqrt{1/\lambda_1^4+1/\lambda_2^4}}{(t_{1/2}^1-t_{1/2}^2)}\right]^2$. Note that since A is not a function of k and D, the constant can be omitted without affecting the final solution.

The Newton-Raphson method for solving the above equations requires calculation of the corresponding second-order derivatives:

$$\frac{\partial^2 C}{\partial k^2} = \frac{12\sigma_e^2}{D^2}\left[\frac{-4c_1}{(k+1)^3} + \frac{(c_3+c_2 D)2(k^3-3k^2-3k-1)}{(k^2+k)^3}\right]$$

$$\frac{\partial^2 C}{\partial D^2} = \frac{24(k-1)\sigma_e^2}{(k+1)k}\left(\frac{3c_1 k}{D^4} + \frac{c_2}{D^3} + \frac{3c_3}{D^4}\right)$$

$$\frac{\partial^2 C}{\partial k \partial D} = -12\sigma_e^2\left[\frac{4c_1}{(k+1)^2 D^3} + \frac{1+2k-k^2}{(k^2+k)^2}\left(\frac{2c_3}{D^3}c_2 D^2\right)\right].$$

Given the preliminary values, (k_0, D_0), the algorithm calculates updated values for k and D according to:

$$\begin{bmatrix} k \\ D \end{bmatrix} = \begin{bmatrix} k_0 \\ D_0 \end{bmatrix} - \begin{bmatrix} \frac{\partial^2 C}{\partial k^2} & \frac{\partial^2 C}{\partial k \partial D} \\ \frac{\partial^2 C}{\partial D \partial k} & \frac{\partial^2 C}{\partial D^2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial C}{\partial k} \\ \frac{\partial C}{\partial D} \end{bmatrix}$$

The algorithm repeatedly updates (k, D) and calculates the above until convergence is obtained.